LEARNING AN INDEPENDENT MULTIMODAL AUDITOR FOR LANGUAGE-GROUNDED SEGMENTATION*

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-grounded segmentation models are increasingly used in multimodal systems, but they can produce confidently incorrect masks, including masks that select the wrong object entirely. This creates a distinct verification problem at deployment: given the image, referring expression, and a candidate mask, determine whether it covers the intended object accurately without seeing the ground-truth mask. A visually coherent segmentation can still be semantically wrong, so quality assessment must check the image–text–mask correspondence as well as mask geometry. Although much work improves mask generation, independent language-conditioned assessment of an already predicted mask remains underexplored. We introduce an independent multimodal large language model (MLLM)-based auditor that verifies predicted masks using only the image, referring expression, and mask itself, without access to ground truth or segmentation model-specific information. We formulate mask verification as estimating the quality of a predicted mask and show that language-conditioned attention provides a useful grounding signal to improve verification. The proposed auditor achieves a better balance between retaining correct masks and rejecting incorrect ones, outperforming existing mask-quality signals and learning-based baselines. It also generalizes across segmentation models and datasets, including challenging reasoning-based referring expressions. These results show that separating segmentation generation from verification provides an effective reliability layer for language-grounded vision systems.

1 INTRODUCTION

Language-grounded segmentation connects natural language with pixel-level perception. Recent approaches include MLLM-based frameworks such as LISA (Lai et al., 2024), GLaMM (Rasheed et al., 2024), and X-SAM (Wang et al., 2026), as well as instruction-following models such as SAM3-I (Li et al., 2026). Their predictions are increasingly used in multimodal and agentic systems, including robotic manipulation in which language-grounded masks guide downstream object localization and action execution (Cho et al., 2026; Huang et al., 2025). In these systems, accepting an incorrect mask can lead to erroneous downstream actions, while rejecting a correct mask can interrupt an otherwise valid action.

Language-grounded segmentation can fail in two fundamentally different ways. A model may identify the correct referent but produce an incomplete or spatially inaccurate mask, resulting in a *geometric error*. Alternatively, it may produce a visually plausible mask for the wrong object, resulting in a *semantic grounding error*. To estimate how likely such failures can occur in practice, we evaluated SAM3-I. Among 10,834 SAM3-I predictions on RefCOCO val, 14.7% have IoU below 0.5. Of these incorrect masks, 82.2% have near-zero overlap with ground truth (IoU < 0.1), while 17.8% have partial overlap ($0.1 \leq$ IoU < 0.5). Near-zero overlap may result from selecting the wrong object or from severe geometric errors. Reliable verification must therefore assess both whether the correct object was selected and whether it was accurately segmented. Fig. 1 illustrates both failure modes alongside a correct prediction.

Many segmentation models provide a confidence or mask-quality score alongside each prediction, typically to rank or filter their own candidate masks. However, such scores are produced within

*Code will be released upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

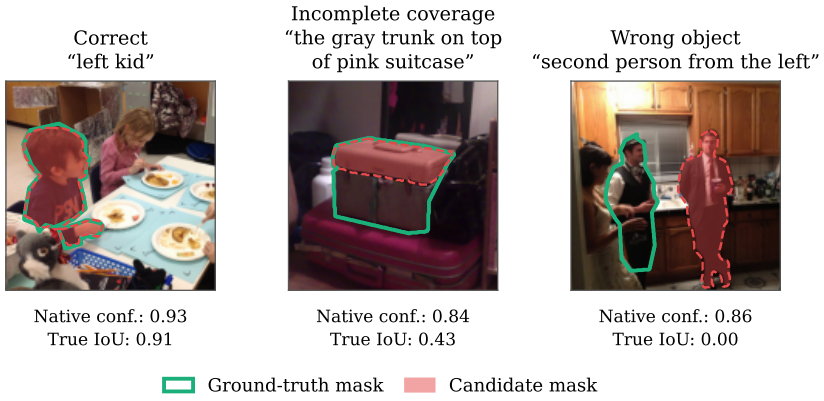


Figure 1: **Failure modes in language-grounded segmentation.** Examples of correct segmentation, geometric error (incomplete coverage), and semantic grounding error (wrong object). Green outlines show ground truth, and red regions show SAM3-I predictions. Native confidence is the SAM3-I's confidence score, while true IoU measures overlap with ground truth. Both incorrect masks receive high confidence despite low IoU.

the same prediction pipeline and do not provide independent verification of mask correctness. Prior work has explored learned mask-IoU prediction (Huang et al., 2019), reference-free quality estimation (Valindria et al., 2017), and in-context variants (Cosarinsky et al., 2025). These approaches largely assess segmentation quality without explicitly verifying consistency with a natural-language referring expression. Recent work has also explored multimodal auditing for language-referred audio-visual segmentation (Zhou et al., 2026). Yet, for image-based language-grounded segmentation, **it remains unclear whether an independent auditor trained on one mask generator can reliably verify predictions from unseen generators and generalize across image and language distributions.**

To address this gap, we introduce an *independent auditor* that verifies language-grounded segmentation masks. Given an image, a referring expression, and a candidate mask, the auditor estimates the quality of the mask using only these inputs, without access to the ground-truth mask or internal information from the mask generator at inference time. This explicitly separates generation from verification: one model produces the segmentation, while an independent model determines whether that prediction should be trusted. The resulting quality estimate can be used both to assess mask accuracy and to decide whether a prediction should be accepted or rejected before being passed downstream.

Effective auditing requires jointly assessing whether the candidate mask corresponds to the object described by the language and whether it adequately covers that object. We therefore fine-tune a MLLM to reason over the original image, the referring expression, and a visualization of the candidate mask. We further investigate whether the MLLM's internal representations already contain useful grounding information. Based on its language-conditioned attention, we derive an attention-mass grounding signal and make this information explicit to the auditor as an additional reliability prior.

Our experiments show that the resulting auditor generalizes beyond the mask generator used to construct its training data. Although trained only on SAM3-I-generated masks from RefCOCO (Yu et al., 2016), it transfers without adaptation to masks produced by X-SAM and outperforms native generator confidence and other competitive learned and zero-shot baselines. We further evaluate on ReasonSeg (Lai et al., 2024), which introduces a different image distribution and more implicit, reasoning-oriented referring expressions. Across these settings, the auditor rejects substantially more bad masks while retaining most good predictions. Finally, to test whether the auditor really understands the relation between the referring expression and the mask rather than judging the mask by appearance alone, we keep the image and candidate mask fixed and change only the referring expression to refer to a different object in the same image. The auditor's quality estimate changes accordingly, reflecting whether the fixed mask matches the newly specified object.

Our main contributions are:

- We formulate the problem of *independent auditing* for language-grounded segmentation, where a verifier must assess both geometric mask quality and semantic consistency with the referring expression, without access to ground-truth masks or internal information from the mask generator at inference time.
- We introduce a generator-independent MLLM auditor that predicts the quality of a candidate mask directly from the image, referring expression, and candidate mask. We further study the MLLM's language-conditioned attention and derive an attention-mass grounding signal that provides an additional view of whether the candidate region is aligned with the referring expression.
- We show that the auditor generalizes without retraining: trained only on SAM3-I masks from RefCOCO, it transfers to the unseen X-SAM generator and remains effective under the harder distribution shift of ReasonSeg. Across these settings, it substantially reduces bad-mask acceptance while retaining most good masks. A controlled text-swap experiment further shows that the auditor's estimates depend on the referring expression rather than mask appearance alone.

2 Problem Formulation

Let I denote an image, T a referring expression, and G a language-conditioned segmentation model that produces a candidate mask $M = G(I, T)$. Let M^{GT} denote the ground-truth mask corresponding to T. We define the true quality of the candidate mask as

$$q = \mathrm{IoU}(M, M^{\mathrm{GT}}) = \frac{|M \cap M^{\mathrm{GT}}|}{|M \cup M^{\mathrm{GT}}|}.$$

Our goal is to learn an independent auditor that predicts this quality, $\hat{q} = A(I, T, M)$. While computing the true IoU requires ground truth, the auditor estimates it using only the image, referring expression, and candidate mask, without accessing the generator's internal representations or confidence scores.

For downstream applications, the predicted quality can serve as a continuous reliability estimate or be thresholded to decide whether a candidate mask should be accepted for further processing. We adopt an IoU threshold of 0.5, an established evaluation criterion in object detection and instance segmentation (Everingham et al., 2010; Lin et al., 2014). Accordingly, we define $y = \mathbf{1}_{\{q \geq 0.5\}}$ and $\hat{y} = \mathbf{1}_{\{\hat{q} \geq 0.5\}}$. Here, y indicates whether the generator's candidate mask is truly good, while $\hat{y}$ represents the auditor's accept–reject decision. Detailed evaluation metrics are defined in Section 4.

3 The Independent Auditor

3.1 Multimodal Auditor

We instantiate the auditor using Qwen2.5-VL-7B-Instruct Bai et al., 2025 with LoRA fine-tuning. For each candidate mask, the MLLM receives the original image, the referring expression, and a second view of the image with the candidate mask overlaid. Providing both images allows the auditor to examine the referred object and assess the spatial extent of the proposed segmentation.

We formulate mask-quality estimation as a next-token prediction task. During training, the prompt asks the MLLM to estimate the IoU of the highlighted mask with the object described by the referring expression. We compute the true IoU using the candidate mask and its ground-truth mask and provide this value as the target textual response. The model is fine-tuned using the standard autoregressive language-modeling objective to predict the tokens representing this numerical value. At inference time, the MLLM generates an IoU estimate from the image, referring expression, and candidate-mask visualization, without access to ground truth or generator-specific information. Fig. 2 illustrates the auditing framework and its attention-prior variant.

3.2 Training Data Construction with Diverse Mask Errors

The candidate masks generated by SAM3-I on RefCOCO are strongly skewed toward high-quality predictions. The full training pool contains 118,624 masks, of which 102,761 have IoU ≥ 0.5 and

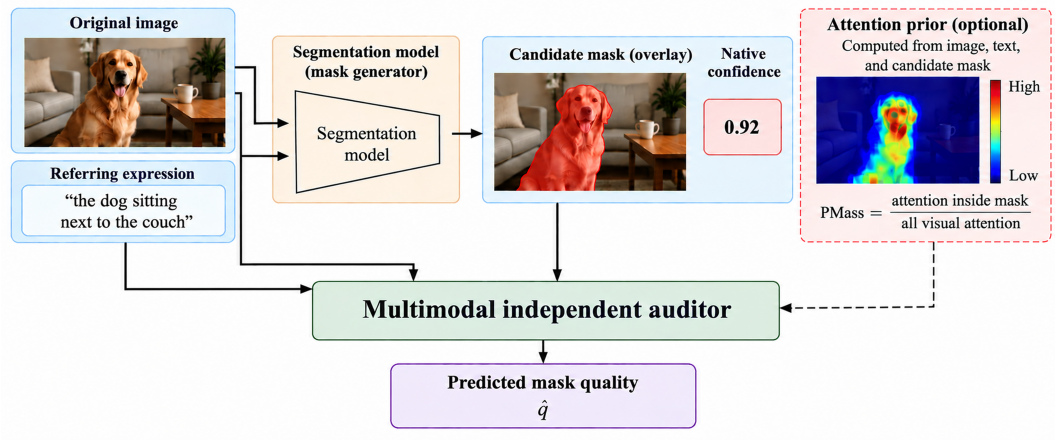


Figure 2: **Overview of independent mask-quality auditing.** A segmentation model generates a candidate mask from an image and referring expression. Our independent multimodal auditor predicts mask quality using the original image, referring expression, and candidate-mask overlay, without access to the generator's native confidence. The attention-prior variant additionally incorporates PMass, derived from language-conditioned visual attention.

only 15,863 have IoU < 0.5. Directly training on this distribution therefore provides comparatively few examples of segmentation failures.

We construct a training set that preserves diversity among good masks while increasing coverage of challenging mask errors. We retain all 15,863 naturally occurring bad masks, sample 51,380 good masks, and generate 15,000 additional challenging bad masks by perturbing high-quality masks. The perturbations include truncation, region removal, region addition or leakage, spatial shifts, erosion, and dilation. Their strength is adjusted so that the resulting masks primarily fall in the challenging IoU range $[0.2, 0.5)$, close to the acceptance boundary.

The resulting training set contains 82,243 examples, with 51,380 good masks and 30,863 bad masks. This construction increases exposure to realistic and near-threshold failures without discarding most of the diversity among high-quality examples. We later compare this training strategy with training on the full dataset.

3.3 Attention-Mass Grounding Signal

Prior work has shown that visual grounding in MLLMs can be concentrated in a small subset of attention heads (Kang et al., 2025). Averaging attention across all heads may therefore dilute localized grounding signals, motivating us to identify whether a grounding head provides complementary information for mask verification.

We define *PMass* as the fraction of language-conditioned visual attention that falls inside the candidate mask. To identify the relevant language tokens, we process the referring expression using spaCy, which assigns part-of-speech tags to individual words. We select content-bearing words, including nouns, proper nouns, adjectives, and verbs, and map them to their corresponding MLLM token positions. Let $\mathcal{Q}(T)$ denote the selected language tokens, $\mathcal{V}$ the set of all visual tokens, and $\mathcal{V}(M)$ the visual tokens corresponding to candidate mask M. For layer ℓ and attention head h, PMass is defined as

$$\mathrm{PMass}_{\ell,h}(M,T) = \frac{1}{|\mathcal{Q}(T)|} \sum_{t \in \mathcal{Q}(T)} \frac{\sum_{i \in \mathcal{V}(M)} a_{t,i}^{(\ell,h)}}{\sum_{i \in \mathcal{V}} a_{t,i}^{(\ell,h)}},$$

where $a_{t,i}^{(\ell,h)}$ denotes the attention from language token t to visual token i. For each selected language token, we calculate the proportion of its visual attention that falls inside the candidate mask and then average these proportions across the selected tokens.

We select an attention head using two disjoint subsets of RefCOCO training data. First, we rank all heads by their average PMass within ground-truth masks and retain the top candidates. For



Figure 3: **Visualizing attention within candidate masks.** Two examples of candidate masks of SAM3-I and their corresponding language-conditioned attention heatmaps. High PMass indicates that most visual attention falls within the candidate mask.

each shortlisted head, we then compute PMass within SAM3-I-generated masks and measure its correlation with the true IoU between these masks and ground truth. We select the head with the highest correlation (layer 16, head 1) for all subsequent experiments. Further details are provided in Appendix A.2. Fig. 3 visualizes attention from the selected head for two SAM3-I candidate masks. In both examples, attention concentrates within the correctly segmented objects, resulting in high PMass scores.

We investigate PMass both as a standalone mask-quality signal and as an additional input to the learned auditor. For the latter, we construct an *attention-prior variant* by explicitly including the scalar PMass value in the MLLM's textual prompt alongside the original image, referring expression, and candidate-mask visualization, as illustrated in Fig. 2. This variant is fine-tuned using the same next-token IoU-prediction objective as the base auditor, allowing us to examine whether explicitly providing internal grounding information improves mask-quality estimation.

4 Experimental Setup

Datasets and mask generators. We consider two mask generators: SAM3-I, an instruction-following extension of SAM 3 (Carion et al., 2026), and X-SAM, an MLLM-based segmentation framework. We train our auditor on curated SAM3-I masks from RefCOCO train, as described in Section 3.2. On RefCOCO val, we evaluate in-domain performance using SAM3-I and cross-generator generalization using the unseen X-SAM. We further evaluate X-SAM-generated masks on ReasonSeg val and test, assessing generalization under simultaneous shifts in mask generator, image distribution, and referring expressions. The referring expressions in ReasonSeg are more implicit and reasoning-oriented than those in RefCOCO. All evaluations use the same auditor without adaptation.

Baselines. We compare against five baselines: the mask generator's native confidence; PMass, our attention-based score; CLIP similarity between the referring expression and the masked image region; zero-shot Qwen2.5-VL-7B-Instruct, prompted to estimate mask IoU without fine-tuning; and a supervised CLIP regressor. The regressor concatenates CLIP embeddings (Radford et al., 2021) of the original image, masked foreground, and referring expression with a mask-area feature, then predicts IoU using a small MLP with frozen CLIP encoders. It is trained on SAM3-I masks from RefCOCO using true IoU as supervision. We additionally evaluate our auditor with attention-prior variant described in Section 3.3.

Evaluation Metrics. We report mean absolute error (MAE) and Pearson correlation with ground truth IoU for continuous quality estimation. For mask acceptance, we report recall, specificity, and macro F1. Recall measures the fraction of good masks correctly accepted ($\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$), while specificity measures the fraction of bad masks correctly rejected ($\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$). Ideally, we want high recall and high specificity, which is reflected in macro F1 score. We also report the percentage of bad masks incorrectly accepted and AUROC for threshold-free discrimination.

5 Results

We first evaluate in-domain performance, followed by generalization to an unseen mask generator and to a different dataset. We then examine how these improvements translate into rejecting incorrect masks.

Table 1: **In-domain and cross-generator evaluation on RefCOCO val.** (a) In-domain evaluation on SAM3-I masks. (b) Cross-generator evaluation on X-SAM masks. The auditor is trained exclusively on curated SAM3-I masks from RefCOCO train and is evaluated on X-SAM without retraining or adaptation. Best MAE, Pearson correlation, and macro F1 are bold.

Method	MAE	Pearson r	Recall	Spec.	Macro F1
(a) SAM3-I: In-domain evaluation					
Native confidence	0.156	0.492	1.000	0.000	0.463
Attention-PMass	0.369	0.375	0.510	0.873	0.510
CLIP similarity	0.613	0.049	0.000	1.000	0.121
Zero-shot Qwen2.5-VL	0.199	0.032	0.965	0.041	0.488
Supervised CLIP regressor	0.171	0.436	0.982	0.154	0.585
Auditor (Ours)	0.091	0.776	0.936	0.823	0.847
Auditor + attention prior (Ours)	**0.080**	**0.808**	0.958	0.865	**0.890**
(b) X-SAM: Cross-generator evaluation					
Native confidence	0.135	0.365	0.999	0.029	0.513
Attention-PMass	0.407	0.255	0.499	0.827	0.411
CLIP similarity	0.674	0.050	0.000	1.000	0.053
Zero-shot Qwen2.5-VL	0.146	0.051	0.966	0.058	0.513
Supervised CLIP regressor	0.142	0.290	0.973	0.137	0.566
Auditor (Ours)	0.087	0.538	0.937	0.647	0.717
Auditor + attention prior (Ours)	**0.079**	**0.568**	0.953	0.698	**0.763**

In-domain performance. Table 1 (a) evaluates SAM3-I masks on RefCOCO val. Most baselines exhibit a pronounced trade-off between recall and specificity: native confidence accepts every incorrect mask, while Attention-PMass rejects most incorrect masks but also discards many correct ones. Zero-shot Qwen2.5-VL and the supervised CLIP regressor similarly favor recall over specificity. In contrast, our auditor achieves recall of 0.936 and specificity of 0.823, yielding a macro F1 of 0.847, compared with 0.585 for the strongest baseline. **Adding the attention prior further improves macro F1 to 0.890**, while also achieving the lowest MAE (0.080) and highest Pearson correlation (0.808).

Cross-generator generalization. We next evaluate the same auditor on X-SAM masks from RefCOCO val, without retraining or exposure to X-SAM during training Table 1(b). **Our auditor generalizes to the unseen generator**, achieving a macro F1 of 0.717, compared with 0.566 for the strongest baseline. It retains 93.7% of correct masks while rejecting 64.7% of incorrect ones, compared with only 2.9% rejected by native confidence. The attention prior further improves macro F1 to 0.763.

Table 2: **Generalization to ReasonSeg.** Performance on X-SAM masks from ReasonSeg val and test. The auditor is trained only on SAM3-I masks from RefCOCO. Full results are in the appendix.

	Val		Test	
Method	Spec.	Macro F1	Spec.	Macro F1
Native confidence	0.310	0.595	0.378	0.577
Attention-PMass	0.714	0.564	0.704	0.582
Zero-shot Qwen2.5-VL	0.155	0.500	0.075	0.415
Supervised CLIP regressor	0.429	0.441	0.503	0.500
Auditor (Ours)	0.750	0.699	0.818	0.705
Auditor + attention prior (Ours)	0.774	**0.733**	0.845	**0.739**

Generalization beyond training data We further evaluate on ReasonSeg, introducing additional shifts in image distribution and referring expressions alongside the unseen X-SAM generator. As shown in Table 2, our auditor outperforms all baselines on both validation and test, achieving macro F1 scores of 0.699 and 0.705, respectively. On test, specificity increases from 0.378 for native

confidence to 0.818 for our auditor, at the cost of lower recall. **Adding the attention prior further improves macro F1 to** 0.733 **on validation and** 0.739 **on test**, with test specificity reaching 0.845.

Bad-mask rejection. Table 3 shows how effectively our auditor rejects incorrect masks. Compared with native confidence, it reduces bad-mask acceptance by 82.3 percentage points on SAM3-I/RefCOCO, 61.8 on X-SAM/RefCOCO, and 44.0 on X-SAM/ReasonSeg. Adding the attention prior further reduces bad-mask acceptance on all three benchmarks.

Table 3: **Bad-mask acceptance (%; lower is better).** Gain is the reduction achieved by our auditor over native confidence, in percentage points.

Generator/Benchmark	Bad (%)	Native	Auditor	+ Attn.Prior	Gain (pp)
SAM3-I / RefCOCO	14.69	100.0	17.7	**13.6**	82.3
X-SAM / RefCOCO	5.80	97.1	35.3	**30.2**	61.8
X-SAM / ReasonSeg	46.47	62.2	18.2	**15.5**	44.0

Threshold-free discrimination. Table 4 evaluates discrimination across all possible decision thresholds. Our auditor achieves AUROC scores of 0.897, 0.797, and 0.755 on SAM3-I/RefCOCO, X-SAM/RefCOCO, and ReasonSeg test, respectively. Adding the attention prior further improves AUROC to 0.935, 0.827, and 0.775, respectively, demonstrating improved discrimination beyond a single acceptance threshold.

Table 4: **Threshold-free discrimination of good and bad masks.** AUROC across in-domain, cross-generator, and cross-dataset evaluations. Higher is better.

Method	SAM3-I	X-SAM	ReasonSeg test
Native confidence	0.809	0.725	0.652
Attention-PMass	0.785	0.732	0.668
Zero-shot Qwen2.5-VL	0.508	0.502	0.517
Supervised CLIP regressor	0.787	0.739	0.509
Auditor (Ours)	0.897	0.797	0.755
Auditor + attention prior (Ours)	**0.935**	**0.827**	**0.775**

Overall, while baselines often favor either recall or specificity, our auditor achieves a better balance and substantially reduces bad-mask acceptance. Attention-PMass, derived from the MLLM's internal attention, also provides a competitive standalone signal, with performance comparable to native confidence. Our auditor generalizes to unseen generators and datasets with sophisticated referring expressions without adaptation, while the attention prior further improves performance across all three settings.

Qualitative examples. Fig. 4 shows four incorrect masks assigned high confidence by the generator SAM3-I (0.90–0.95), despite having almost no overlap with ground truth. In all four cases, our auditor predicts an IoU of 0.00 and correctly rejects the masks. These examples illustrate how independent verification can detect incorrect object selection that native confidence fails to identify.

6 Analysis

We examine how PMass relates to mask quality, how our curated training data affects performance, and whether the auditor genuinely uses the referring expression when evaluating a mask.

Comparison of mask-quality signals. Fig. 5 compares native confidence, PMass, and our auditor with attention-prior against true mask IoU from RefCOCO val data. For low-IoU masks, native confidence remains high, with scores concentrated around 0.8, which nicely demonstrates the overconfidence of SAM3-I in bad masks. PMass assigns lower but more variable scores and generally increases with IoU, providing a proxy for how well the mask matches the object described by the text. In contrast, the attention-prior auditor produces quality estimates that better follow the true IoU

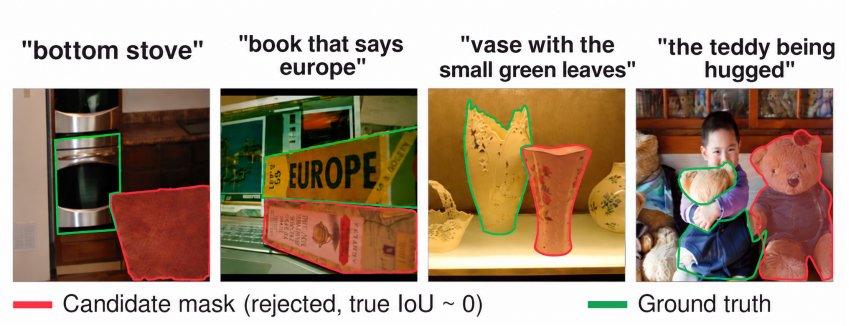


Figure 4: **High-confidence incorrect masks rejected by our auditor.** Despite high native confidence (0.90–0.95), all four masks have near-zero true IoU. The auditor predicts an IoU of 0.00 for each and rejects all four. Red indicates candidate masks; green indicates ground truth.

distribution, assigning lower scores to poor masks and higher scores to accurate masks. This results in a stronger correlation with true IoU ($r = 0.808$).

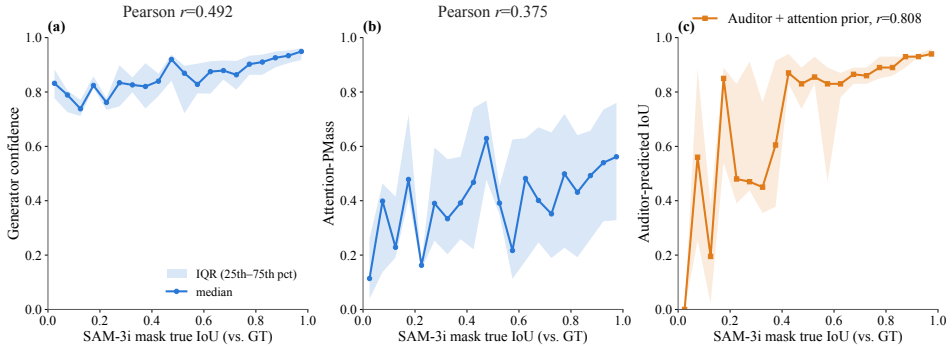


Figure 5: **Comparison of mask-quality signals with true IoU.** Native confidence remains high even for low-quality masks, while PMass provides a complementary attention-based grounding signal. The attention-prior auditor produces quality estimates that more closely track true IoU, achieving higher Pearson correlation ($r = 0.808$). Lines show binned medians and shaded regions show interquartile ranges.

Effect of training-data construction. Table 5 compares the same auditor trained on the full SAM3-I mask pool from RefCOCO train and our curated training set. On both in-domain SAM3-I and cross-generator X-SAM evaluations, full-data training achieves higher recall, while curated training achieves higher specificity, with only small differences in macro F1 (0.022 and 0.017, respectively). On ReasonSeg test, curated training substantially improves specificity and increases macro F1 from 0.646 to 0.705, although recall decreases. Overall, curated training makes the auditor more effective at rejecting incorrect masks, at the cost of rejecting some additional correct masks.

Table 5: **Effect of training data on auditor performance.** Both auditors are trained exclusively on SAM3-I masks from RefCOCO train. Full uses the original training pool; Curated uses our training set enriched with incorrect masks.

Benchmark	Training	Recall	Spec.	Macro F1
SAM3-I / RefCOCO	Full	0.986	0.681	**0.869**
	Curated	0.936	**0.823**	0.847
X-SAM / RefCOCO	Full	0.984	0.416	**0.734**
	Curated	0.937	**0.647**	0.717
X-SAM / ReasonSeg test	Full	0.783	0.514	0.646
	Curated	0.608	**0.818**	**0.705**

Does the auditor use the referring expression? To test whether the auditor uses language when evaluating a mask, we construct 1,500 text-swap pairs from RefCOCO. For each pair, we keep the image and candidate mask unchanged but replace the referring expression with one describing a different object in the same image. We then recompute the true IoU for the newly referred object.

As shown in Table 6, the auditor's mean predicted IoU drops from 0.801 to 0.081, closely following the decrease in true IoU from 0.753 to 0.041. Moreover, its predictions for the swapped expressions correlate with the new object's true IoU ($r = 0.669$), rather than the original object's IoU ($r = -0.356$). **These results show that the auditor uses the referring expression to assess whether the mask covers the specified object, rather than relying only on the image and candidate mask.**

Table 6: **Text-swap evaluation on 1,500 pairs.** The image and candidate mask remain fixed while the referring expression changes to describe another object in the same image. Correlations compare the swapped predictions with the true IoU for each referred object.

Metric	Original referent	Swapped referent
Mean predicted IoU	0.801	0.081
Mean true IoU	0.753	0.041
Correlation with swapped predictions (r)	-0.356	$+0.669$

7 Related Work

Language-Grounded Segmentation. Early referring-expression methods (Yu et al., 2016; Hu et al., 2016) and vision-language models (Radford et al., 2021; Kamath et al., 2021; Lüddecke & Ecker, 2022; Subramanian et al., 2022) established text-guided localization. SAM, SegGPT, and SAM 3 advanced general-purpose segmentation (Kirillov et al., 2023; Wang et al., 2023; Carion et al., 2026), while LISA, GLaMM, GSVA, and PSALM integrated multimodal language models with segmentation (Lai et al., 2024; Rasheed et al., 2024; Xia et al., 2024; Zhang et al., 2024). Recent SAM3-I and X-SAM further extend instruction-based and general-purpose segmentation (Li et al., 2026; Wang et al., 2026). Unlike these methods, we focus on independently verifying generated masks.

Segmentation Quality Assessment. Mask Scoring R-CNN predicts mask IoU to improve quality estimation (Huang et al., 2019). Reverse Classification Accuracy and In-Context RCA estimate segmentation quality without ground truth at inference time (Valindria et al., 2017; Cosarinsky et al., 2025). Our auditor additionally evaluates whether a mask matches the object specified by the referring expression. **The distinction is consequential: a mask may delineate an object cleanly while selecting the wrong referent.** We therefore assess quality conditioned on the expression and test whether a verifier trained with masks from one generator can judge masks from another.

Multimodal Auditing and Attention. The closely related MQA-RefAVS framework uses an MLLM to assess language-referred audio-visual segmentation (Zhou et al., 2026). We instead study image-based mask verification across unseen generators and datasets. Prior work has explored attention for segmentation, visual grounding, and vision-language alignment (Wu et al., 2025; Kang et al., 2025; Aflalo et al., 2025). We extend this direction by evaluating attention mass as a mask-quality signal and incorporating it as an explicit prior into an independent auditor.

8 Conclusion

We introduced an independent MLLM-based auditor for assessing the reliability of language-grounded segmentation masks. The auditor achieves a better balance between preserving correct masks and rejecting unreliable ones, outperforming the original mask-generator mask-quality signals and learning-based baselines. Importantly, it generalizes beyond the segmentation models and datasets used during training, demonstrating robust verification across different evaluation settings. Incorporating language-conditioned attention as an attention prior further improves verification by providing an additional grounding signal. Our results suggest that reliability in language-grounded vision systems can benefit from separating the role of producing predictions from the role of verifying them.

AI Assistance Disclosure

We used large language models to assist with language editing and improving the clarity of the manuscript. All scientific ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors.

References

Estelle Aflalo, Gabriela Ben Melech Stan, Tiep Le, Man Luo, Shachar Rosenman, Sayak Paul, Shao-Yen Tseng, and Vasudev Lal. Fivl: A framework for improved vision-language alignment through the lens of training, evaluation and explainability, 2025. URL https://arxiv.org/abs/2412.14672.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, and Min-Hung Chen. Spatialclaw: Rethinking action interface for agentic spatial reasoning, 2026. URL https://arxiv.org/abs/2606.13673.

Matias Cosarinsky, Ramiro Billot, Lucas Mansilla, Gabriel Jimenez, Nicolas Gaggion, Guanghui Fu, and Enzo Ferrante. In-context reverse classification accuracy: Efficient estimation of segmentation quality without ground-truth. *arXiv preprint arXiv:2503.04522*, 2025.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pp. 108–124, 2016. doi: 10.1007/978-3-319-46448-0_7.

Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. Roboground: Robotic manipulation with grounded vision-language priors, 2025. URL https://arxiv.org/abs/2504.21530.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Laura Rolland, David Schmidt, Zhengyan Li, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Jingjing Li, Yue Feng, Yuchen Guo, Jincai Huang, Wei Ji, Qi Bi, Yongri Piao, Miao Zhang, Xiaoqi Zhao, Qiang Chen, Shihao Zou, Huchuan Lu, and Li Cheng. Sam3-i: Segment anything with instructions, 2026. URL https://arxiv.org/abs/2512.04585.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096, 2022. doi: 10.1109/CVPR52688.2022.00695.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. ReCLIP: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5198–5215, 2022. doi: 10.18653/v1/2022.acl-long.357.

Vanya V. Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O. Aboagye, Andrea G. Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*, 36(8):1597–1606, 2017.

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-sam: From segment anything to any segmentation, 2026. URL https://arxiv.org/abs/2508.04655.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1130–1140, 2023.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24710–24721, 2025.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models, 2024. URL https://arxiv.org/abs/2312.10103.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Jinxing Zhou, Yanghao Zhou, Yaoting Wang, Zongyan Han, Jiaqi Ma, Henghui Ding, Rao Muhammad Anwer, and Hisham Cholakkal. Audit after segmentation: Reference-free mask quality assessment for language-referred audio-visual segmentation. *arXiv preprint arXiv:2602.03892*, 2026.

A Appendix

A.1 Full ReasonSeg results.

Table 7 provides additional quality-estimation metrics for ReasonSeg validation and test. On validation, the attention-prior auditor improves Pearson correlation from 0.518 to 0.545 and macro F1

from 0.699 to 0.733, with nearly identical MAE. On test, it improves specificity from 0.818 to 0.845 and macro F1 from 0.705 to 0.739, although the base auditor achieves slightly lower MAE and higher correlation. These results show that the attention prior improves mask acceptance decisions without uniformly improving continuous IoU estimation.

Table 7: **Full ReasonSeg evaluation.** All masks are generated by X-SAM. The auditors are trained only on curated SAM3-I masks from RefCOCO train. Best MAE, Pearson correlation, and macro F1 are bold.

Method	MAE	Pearson r	Spec.	Macro F1
(a) ReasonSeg validation				
Native confidence	0.327	0.319	0.310	0.595
Attention-PMass	0.365	0.227	0.714	0.564
CLIP similarity	0.443	0.061	1.000	0.296
Zero-shot Qwen2.5-VL	0.342	0.200	0.155	0.500
Supervised CLIP regressor	0.388	-0.109	0.429	0.441
Auditor (Ours)	**0.254**	0.518	0.750	0.699
Auditor + attention prior (Ours)	0.255	**0.545**	0.774	**0.733**
(b) ReasonSeg test				
Native confidence	0.338	0.303	0.378	0.577
Attention-PMass	0.336	0.316	0.704	0.582
CLIP similarity	0.412	0.191	1.000	0.320
Zero-shot Qwen2.5-VL	0.387	0.088	0.075	0.415
Supervised CLIP regressor	0.376	-0.003	0.503	0.500
Auditor (Ours)	**0.263**	**0.515**	0.818	0.705
Auditor + attention prior (Ours)	0.270	0.496	0.845	**0.739**

A.2 Additional Experimental Details

Training Configuration. We fine-tune Qwen2.5-VL-7B-Instruct using LoRA with rank 64, alpha 128, and dropout 0.05. LoRA is applied to the language model's linear layers. The vision encoder remains frozen, while the vision–language projector is fully fine-tuned.

Training uses the standard next-token prediction objective, AdamW, a linear learning-rate schedule, a warmup ratio of 0.03, and bf16 precision. We train for one epoch using two NVIDIA A100 80GB GPUs, with a per-device batch size of 8 and gradient accumulation of 8, giving an effective batch size of 128. Images are processed using Qwen2.5-VL's default dynamic-resolution processor.

Training-Set Construction. Our original training pool contains 118,624 SAM3-I predictions from RefCOCO train: 102,761 good masks (IoU ≥ 0.5) and 15,863 bad masks (IoU < 0.5). True IoU is computed directly from the predicted and ground-truth masks using `pycocotools`.

We construct a curated set of 82,243 examples, consisting of 51,380 randomly selected good masks, all 15,863 natural bad masks, and 15,000 synthetic bad masks.

Synthetic masks are generated by perturbing correct SAM3-I predictions. We apply six types of corruption: truncation, region removal, spatial shifts, region addition, erosion, and dilation. Original masks used to generate perturbations are excluded from the sampled good-mask records.

The attention-prior auditor uses the same curated training examples, with an additional PMass score included in each prompt. For synthetic examples, PMass is recomputed using the perturbed mask.

Inference and Evaluation. All auditor variants and the zero-shot Qwen2.5-VL baseline use the same prompt format. The model receives the original image, the candidate mask overlay, and the referring expression, and is asked to estimate the IoU between the candidate mask and the unseen ground-truth mask.

The model is instructed that errors may include missing object regions, including extra regions, or selecting the wrong object. It outputs a JSON object containing the estimated IoU and confidence score.

Training targets use the true IoU rounded to two decimal places and a fixed confidence of 0.9. At inference time, we extract the predicted IoU from the response and clip it to $[0, 1]$.

A.2.1 ATTENTION-PMASS DETAILS

Attention Head Selection. We select the PMass attention head in two phases using two disjoint RefCOCO training splits. First, we rank all 784 attention heads by their average PMass computed inside ground-truth masks. This phase does not use generated masks or IoU information. Layer 16, head 1 achieves the highest average PMass (0.486).

In the second phase, we evaluate the shortlisted heads using SAM3-I-generated candidate masks. For each head, we compute PMass within the candidate mask and measure its correlation with the true IoU. Layer 16, head 1 achieves the highest correlation (Pearson $r = 0.313$, Spearman $\rho = 0.271$) and is used in all experiments.

We also compare different language-token aggregation strategies. Averaging attention over content words (nouns, proper nouns, adjectives, and verbs) provides the best performance compared with whole-prompt or whole-expression aggregation. Using additional heads does not improve performance, so we use a single attention head for simplicity.

PMass Computation. PMass is computed using the original, frozen Qwen2.5-VL-7B-Instruct model. We use spaCy to select content-bearing words from the referring expression, including nouns, proper nouns, adjectives, and verbs.

For each selected language token, we calculate the fraction of its visual attention that falls inside the candidate mask and average these fractions across the selected tokens. A visual token is considered inside the mask when more than 50% of its corresponding image patch is covered.

PMass is calculated separately for each candidate mask without using ground truth at inference time.

A.2.2 BASELINE IMPLEMENTATION

Native confidence. We use the confidence score directly provided by SAM3-I or X-SAM, without modification.

Attention-PMass. We use PMass from layer 16, head 1 of the frozen Qwen2.5-VL model as a standalone mask-quality score, without fine-tuning.

CLIP similarity. We compute cosine similarity between the referring expression and the masked foreground using CLIP ViT-L/14. Pixels outside the candidate mask are replaced with gray. This baseline requires no training.

Zero-shot Qwen2.5-VL. We use the original Qwen2.5-VL-7B-Instruct without fine-tuning, with the same two-image prompt and response-parsing procedure as our auditor.

Supervised CLIP regressor. We extract frozen CLIP ViT-L/14 embeddings from the original image, masked foreground, and referring expression. These features, together with the mask-area fraction, are passed to a two-layer MLP with a hidden dimension of 512, GELU activation, dropout of 0.1, and a sigmoid output.

Only the MLP is trained, using the original $118{,}624$ SAM3-I training examples for two epochs. We use Adam with batch size 96, learning rate 10^{-3}, weight decay 10^{-4}, and Smooth-L1 loss.